

MIMCA: multiple imputation for categorical variables with multiple correspondence analysis

Vincent Audigier¹ · François Husson¹ · Julie Josse¹

Received: 21 June 2015 / Accepted: 30 January 2016 / Published online: 11 February 2016
© Springer Science+Business Media New York 2016

Abstract We propose a multiple imputation method to deal with incomplete categorical data. This method imputes the missing entries using the principal component method dedicated to categorical data: multiple correspondence analysis (MCA). The uncertainty concerning the parameters of the imputation model is reflected using a non-parametric bootstrap. Multiple imputation using MCA (MIMCA) requires estimating a small number of parameters due to the dimensionality reduction property of MCA. It allows the user to impute a large range of data sets. In particular, a high number of categories per variable, a high number of variables or a small number of individuals are not an issue for MIMCA. Through a simulation study based on real data sets, the method is assessed and compared to the reference methods (multiple imputation using the loglinear model, multiple imputation by logistic regressions) as well to the latest works on the topic (multiple imputation by random forests or by the Dirichlet process mixture of products of multinomial distributions model). The proposed method provides a good point estimate of the parameters of the analysis model considered, such as the coefficients of a main effects logistic regression model, and a reliable estimate of the variability of the estimators. In addition, MIMCA has the great advantage that it is substantially less time consuming on data sets of high dimensions than the other multiple imputation methods.

Keywords Missing values · Categorical data · Multiple imputation · Multiple correspondence analysis · Bootstrap

Mathematics Subject Classification 62H25 · 6207 · 62F40

1 Introduction

Data sets with categorical variables are ubiquitous in many fields such as social sciences, where surveys are conducted through multiple-choice questions. Whatever the field, missing values frequently occur and are a key problem in statistical practice since most of statistical methods cannot be applied directly on incomplete data. In this paper, we pursue the aim of achieving inference for a parameter of a statistical analysis, denoted ψ , from an incomplete data set. It means to get a point estimate and an estimate of its variance in a missing values framework.

Under the classical missing at random mechanism (MAR) assumption for the incomplete data (Schafer 1997), two main strategies are available to deal with missing values. The first one consists in adapting the statistical analysis so that it can be applied on an incomplete data set. For instance, the maximum likelihood (ML) estimators of ψ can be obtained from incomplete data using an Expectation-Maximization (EM) algorithm (Dempster et al. 1977) and their standard error can be estimated using a Supplemented Expectation-Maximization algorithm (Meng and Rubin 1991). The ML approach is suitable, but often difficult to establish (Allison 2012) and tailored to a specific statistical method.

That is why the second strategy namely multiple imputation (MI) (Rubin 1987; Little and Rubin 1987) seems to take the lead. The principle of MI is to replace the missing values by plausible values and repeat it M times in order to obtain

✉ Vincent Audigier
audigier@agrocampus-ouest.fr

François Husson
husson@agrocampus-ouest.fr

Julie Josse
josse@agrocampus-ouest.fr

¹ Applied Mathematics Department, Agrocampus Ouest,
65 rue de Saint-Brieuc, 35042 Rennes Cedex, France

M imputed data sets. The imputed values are generated from an *imputation model* using M different parameters. This set of parameters reflects the uncertainty on the parameters used to perform the imputation. Then, MI consists in estimating the parameters ψ of the statistical method (called *analysis model*) on each imputed data set. Note that several analysis models can be applied to a same multiply imputed data set, e.g. a logistic regression, a chi-square statistics, a proportion, etc. Lastly, the $(\hat{\psi}_m)_{1 \leq m \leq M}$ estimates of the parameters are pooled to provide a unique estimation for ψ and for its associated variability using Rubin's rules (Rubin 1987). This ensures that the variance of the estimator appropriately takes into account the supplement variability due to missing values.

To get valid inferences for a large variety of analysis models, a desirable property of a MI method is to take into account as many associations between variables as possible (Schafer 2003). The aim is to have an imputation model at least as general as the analysis models. For instance, a proportion can be estimated after using an imputation model not respecting the relationships between variables, but these should be taken into account for application of a logistic regression with main effects or calculation of chi-square statistics.

Suggesting a MI method for categorical data is a very challenging task. Indeed, the imputation models suffer from estimation issues as the number of parameters quickly grows with large number of categories and variables. Ideally, one would like an imputation model which requires a moderate number of parameters, while preserving as much as possible the relationships between variables.

In this paper, we develop a MI method based on multiple correspondence analysis (MCA) named MIMCA for *multiple imputation with multiple correspondence analysis*. MCA is the counterpart of principal component analysis (PCA) for categorical data. Principal component methods are often used to sum up the similarities between the individuals and the relationships between variables using a small number of synthetic variables (principal components) and synthetic observations (loadings). These methods reduce the dimensionality of the data while keeping as much as possible the information of the data. This property is particularly relevant to analyse high dimensional categorical data, but especially appealing to perform MI. Indeed, MCA proposes an attractive trade-off between complexity of the model and preservation of the data structure.

The remainder of this paper is organized as follows. We briefly present in Sect. 2 the advantages and drawbacks of the standard MI methods to deal with categorical data: MI using the loglinear model (Schafer 1997), MI using logistic regressions (Van Buuren 2012). We also quickly describe recent propositions which tackle some of the issues of the former methods such as MI using random forests (Doove et al. 2014) and MI using the latent class model (Si and Reiter 2013). MI using the normal distribution (King et al. 2001) will be dis-

cussed as well to highlight the differences with MIMCA. In Sect. 3, we describe our method and give its properties. In Sect. 4, a simulation study based on real data sets evaluates the novel method and compares its performances to the other main MI methods. All our results are reproducible and the method is available in the package missMDA (Husson and Josse 2015) of the open-source R software (R Core Team 2014). The proposed methodology is detailed on a data set with the R software in Sect. 5.

2 Multiple imputation methods for categorical data

Hereinafter, matrices and vectors will be in bold text, whereas sets of random variables or single random variables will not. Matrices will be in capital letters, whereas vectors will be in lower case letters. We denote $\mathbf{X}_{I \times K}$ a data set with I individuals and K variables and \mathbf{T} the corresponding contingency table. We note the observed part of \mathbf{X} by \mathbf{X}_{obs} and the missing part by \mathbf{X}_{miss} , so that $\mathbf{X} = (\mathbf{X}_{obs}, \mathbf{X}_{miss})$. Let q_k denote the number of categories for the variable \mathbf{X}_k and $J = \sum_{k=1}^K q_k$ the total number of categories. We note $\mathbb{P}(X; \theta)$ the distribution of the variables $X = (X_1, \dots, X_K)$ where θ is the set of parameters of the distribution.

2.1 Multiple imputation using a loglinear model

MI using the loglinear model is considered as the gold standard for MI on small incomplete categorical data sets (Vermunt et al. 2008; van der Palm et al. 2014). The loglinear model (Agresti 2013) assumes a multinomial distribution $\mathcal{M}(\theta, I)$ as joint distribution for \mathbf{T} , where $\theta = (\theta_{x_1 \dots x_K})_{x_1 \dots x_K}$ is a vector indicating the probability to observe each event $(X_1 = x_1, \dots, X_K = x_K)$. MI from the loglinear model is achieved by drawing imputed values from this joint distribution. To do so, a Bayesian treatment of this model is used in which a Dirichlet prior is assumed for θ , implying a Dirichlet distribution for the posterior (Schafer 1997). This method is very attractive since it reflects all kind of relationships between variables, which enables applying any analysis model. However, this method is dedicated to data sets with a small number of categories because it requires a number of independent parameters approximately equal to the number of combinations of categories. For example, it corresponds to 9 765 624 independent parameters for a data set with $K = 10$ variables with $q_k = 5$ categories for each of them. This involves *overfitting* issues, *i.e.* the estimation procedure fits well the observed data whereas missing values are badly predicted. The variability associated to the prediction is far too large. Note that this phenomenon is not specific to the missing data framework. For instance, overfitting is well known for the linear regression model. It occurs when the number of independent parameters is large compared to the

number of observations. Regularization is a way to tackle this issue in the regression framework. Concerning the log-linear model, the issue is overcome by adding constraints on θ to limit the number of independent parameters estimated (Schafer 1997, pp. 289–331). A sparser model commonly used is the model of homogeneous associations (Agresti 2013, p. 344) which takes into account two-way associations between variables only (whereas the multinomial model takes into account higher-order associations). However, the number of independent parameters remains equal to the number of pairs of categories that can be quite large (760 for the previous example).

2.2 Multiple imputation using a latent class model

To overcome the limitation of MI using the loglinear model, another MI method based on the latent class model can be used. The latent class model (Agresti 2013, p. 535) is a mixture model based on the assumption that each individual belongs to a latent class from which all variables can be considered as independent. More precisely, let Z denote the latent categorical variable whose values are in $\{1, \dots, L\}$. Let $\theta_Z = (\theta_\ell)_{1 \leq \ell \leq L}$ denote the proportion of the mixture, $\theta_X = (\theta_x^{(\ell)})_{1 \leq \ell \leq L}$ the parameters of the L components of the mixture and let $\theta = (\theta_Z, \theta_X)$ denote the parameters of the mixture. Thus, assuming a multinomial distribution for Z and $X_k|Z$, the joint distribution of the data is written as follows:

$$\mathbb{P}(X = (x_1, \dots, x_K); \theta) = \sum_{\ell=1}^L \left(\theta_\ell \prod_{k=1}^K \theta_{x_k}^{(\ell)} \right) \quad (1)$$

The latent class model requires $L \times (J - K) + (K - 1)$ independent parameters, *i.e.* a number that linearly increases with the number of categories.

Vidotto et al. (2014) reviews in detail different MI methods using a latent class model. One of the latest contributions in this family of methods uses a non-parametric extension of the model namely the Dirichlet process mixture of products of multinomial distributions model (DPMPM) (Dunson and Xing 2009; Si and Reiter 2013). This method uses a fully Bayesian approach in which the number of classes is defined automatically by specifying a stick breaking process (Ishwaran and James 2001) for the mixture probabilities θ_Z . The prior distribution for the parameters of the components of the mixture θ_X is a Dirichlet distribution. Note that this MI method based on the latent class model is not too computationally intensive (Vidotto et al. 2014).

Because the latent class model approximates quite well any kind of relationships between variables, MI using this model enables the use of complex analysis models such as logistic regression with some interaction terms and provides

good estimates of the parameters of the analysis model. Nevertheless, the imputation model implies that given a class, each individual is imputed in the same way, whatever the categories taken. If the class is very homogeneous, all the individuals have the same observed values, and this behaviour makes sense. However, when the number of missing values is high and when the number of variables is high, it is not straightforward to obtain homogeneous classes. It can explain why Vidotto et al. (2014) observed that the MI using the latent class model can lead to biased estimates for the analysis model in such cases.

2.3 Multiple imputation using a multivariate normal distribution

Because MI based on the normal multivariate distribution is a robust method for imputing continuous non-normal data (Schafer 1997), imputation using the multivariate normal model can be seen as an attractive method for imputing categorical variables recoded as dummy variables. The imputed dummy variables are seen as a set of latent continuous variables from which categories can be independently derived. Since the method we propose shares some common points with this method, let us give more details. Let $\mathbf{Z}_{I \times J}$ denote the disjunctive table coding for $\mathbf{X}_{I \times K}$, *i.e.* the set of dummy variables coding for the incomplete matrix. Note that one missing value on \mathbf{x}_k implies q_k missing values on \mathbf{z}_k , the set of dummy variables in \mathbf{Z} coding for the variable \mathbf{x}_k . The following procedure implemented in Honaker et al. (2014), Honaker et al. (2011) enables the MI of a categorical data set using the normal distribution:

- perform a non-parametric bootstrap on \mathbf{Z} : sample the rows of \mathbf{Z} with replacement M times. M incomplete disjunctive tables $(\mathbf{Z}_m^{boot})_{1 \leq m \leq M}$ are obtained;
- estimate the parameters of the normal distribution on each bootstrap replicate: calculate the ML estimators of (μ_m, Σ_m) , the mean and the variance of the normal distribution for the m th bootstrap incomplete replicate, using an EM algorithm;
- create M imputed disjunctive tables: impute \mathbf{Z} from the normal distribution using $(\mu_m, \Sigma_m)_{1 \leq m \leq M}$ and the observed values of \mathbf{Z} . The M imputed disjunctive tables obtained are denoted $(\mathbf{Z}_m)_{1 \leq m \leq M}$. In \mathbf{Z}_m , the observed values are still zeros and ones, whereas the missing values have been replaced by real numbers;
- generate M imputed categorical data sets: from the latent continuous variables contained in $(\mathbf{Z}_m)_{1 \leq m \leq M}$, choose categories for each incomplete individual.

Several ways have been proposed to get the imputed categories from the imputed continuous values. For example,

Allison (2002) recommends to attribute the category corresponding to the highest imputed value, while Bernaards et al. (2007), Demirtas (2009), Yucel et al. (2008) propose some rounding strategies. However, “A single best rounding rule for categorical data has yet to be identified.” (Van Buuren 2012, p. 107). A common strategy proposed by Bernaards et al. (2007) is called *Coin flipping*. It consists in considering the set of imputed continuous values of the q_k dummy variables \mathbf{z}_k as an expectation given the observed values $\theta_k = \mathbb{E} \left[(z_1, \dots, z_{q_k}) | Z_{obs}; \hat{\mu}, \hat{\Sigma} \right]$. Thus, the imputed categories are obtained by randomly drawing one category according to a multinomial distribution $\mathcal{M}(\theta_k, 1)$. Note that θ_k could be suitably modified so that it remains between 0 and 1: the imputed continuous values lower than 0 are replaced by 0, the imputed values larger than 1 are replaced by 1 and the others are unchanged. In such a case, the imputed values are scaled to respect the constraint that the sum is equal to one per variable.

Because imputation under the normal multivariate distribution is based on the estimate of a covariance matrix, the imputation under the normal distribution can detect only two-way associations between categorical variables, which is generally sufficient for the analysis model. The main drawback of the MI using the normal distribution is the number of independent parameters estimated. This number is equal to $\frac{(J-K) \times (J-K+1)}{2} + (J-K)$, so approximately the square of the total number of categories, which quickly leads to overfitting. Moreover, the covariance matrix is not invertible when the number of individuals is lower than $(J-K)$ or when the collinearity between dummy variables is very high (Carpenter and Kenward 2013, p. 191; Audigier et al. 2014a). This can be a serious drawback for categorical data because collinearity between dummy variables frequently occurs. In this case the conditional distributions of the missing values do not exist. To overcome these issues, it is possible to add a ridge term on its diagonal to improve the conditioning of the regression problem.

2.4 Fully conditional specification

Instead of making MI by specifying a joint model for the variables (JM), like the previous MI methods, categorical data can be imputed using a fully conditional specification (FCS) approach (Buuren et al. 2006): for each variable with missing values, an imputation model is defined (i.e. a conditional distribution) and each incomplete variable is sequentially imputed according to this, while reflecting the uncertainty on the model’s parameters. Implicitly, the choices of the conditional distributions determine a joint distribution, in so far as a joint distribution is compatible with these choices (Besag 1974). Typically, the models used for each incomplete variable are some multinomial logistic regressions and

the variability of the models’ parameters is reflected using a Bayesian point of view. The convergence to the joint distribution is obtained by repeating the conditional imputations several times. The procedure is performed M times in parallel to provide M imputed data sets.

FCS is more computationally intensive than JM (Van Buuren 2012; Vermunt et al. 2008). This is not a practical issue when the data set is small, but it becomes so on a data set of high dimensions. In particular, checking the convergence becomes difficult.

The imputation using multinomial logistic regressions on each variable performs quite well, that is why this method is often used as a benchmark to perform comparative studies (van der Palm et al. 2014; Doove et al. 2014; Shah et al. 2014; Si and Reiter 2013). However, the typical issues of logistic regression can affect the MI procedure using this model. Indeed, when separability problems occur (Albert and Anderson 1984), when the number of individuals is smaller than the number of categories (Agresti 2013, p. 195), or when collinearity occurs (Agresti 2013, p. 208), it is not straightforward to get the estimates of the parameters. In addition, when the number of categories becomes large, too many parameters have to be estimated, implying overfitting.

Typically, the logistic regression models constructed to impute each variable are main effects models. Thus, the imputation model captures the two-way associations between variables well (Agresti 2013; van der Palm et al. 2014). Models taking into account interactions can be used, but the choice of these models requires a certain effort by the user. Note that the conditional distributions defined by the logistic regressions with main effects can be deduced from the joint distribution defined by the loglinear model with two-way associations (Agresti 2013, pp. 353–356). Thus, the corresponding MI methods are very close.

FCS using other conditional models than logistic regression can be used (see Van Buuren 2012 for instance). Among those, conditional imputations using random forests have been recently suggested (Doove et al. 2014; Shah et al. 2014). Random forests are non-parametric models based on draws from the observed values. It captures complex relationships between variables giving an accurate prediction of the missing values whatever the structure of the data set is (Stekhoven and Bühlmann 2012). Moreover, the method can be used with any number of individuals, and any number of variables, which can be an interesting property. Thus, FCS using random forest seems to be a promising MI method. However, it is more computationally intensive than the one based on logistic regressions. In addition, because the method is based on draws from the observed values, it can provide poor imputation for rare categories (Audigier et al. 2014b). According to Doove et al. (2014), an imputation of one variable X_k given the others is obtained as follows:

- build a forest of 10 trees:
 - draw 10 bootstrap samples from the individuals without missing value on X_k ;
 - fit one tree on each bootstrap sample: draw randomly a subset of $\sqrt{K-1}$ variables among the $K-1$ explanatory variables for splitting the bootstrap sample at each node. Find the best split at each node according to the given subset of explanatory variables.

Note that the uncertainty due to missing values is taken into account by the use of several trees (the forest) instead of using a unique tree;

- impute missing values according to the forest:
 - for an individual i with a missing value on X_k , gather all the individuals from the predictive leaf of the 10 trees and draw randomly one individual from it.
 - repeat for all individuals with missing values on X_k .

Then, the procedure is performed for each incomplete variable and repeated until convergence. The method is very robust to the number of trees used, as well as to the number of explanatory variables retained. Thus, the default choices for these parameters (10 trees, $\sqrt{K-1}$ explanatory variables) are very suitable in most of the cases.

3 Multiple imputation using multiple correspondence analysis

This section presents a novel MI method for categorical data based on multiple correspondence analysis (Greenacre and Blasius 2006; Lebart et al. 1984), i.e. the principal component method dedicated to categorical data. Like the imputation using the normal distribution, it is a JM method based on the imputation of the disjunctive table. We first introduce MCA as a specific singular value decomposition (SVD) of the data matrix. Then, we describe how to perform this SVD with missing values and how it is used to perform single imputation. Next, we explain how to take into account the uncertainty on the parameters of MCA to get a MI method. Finally, the properties of the method are discussed and the differences with MI using the normal distribution highlighted.

3.1 MCA for complete data

MCA can be seen as the counterpart of PCA for categorical data, whereas PCA is dedicated to continuous data. MCA is a very popular method to describe, summarise and visualise multidimensional data in order to understand the two-way associations between variables as well as the similarities between individuals. It is especially useful for high dimen-

sional data. Standard references included Benzécri (1973), Nishisato (1980), Lebart et al. (1984), Greenacre (1984), Gifi (1981) and Greenacre and Blasius (2006).

MCA is a dimensionality reduction method consisting in searching for a subspace of dimension S providing the best representation of the categorical data in the sense that it maximises the variability of the projected points. Like any principal component method, it boils down to performing a SVD of the data matrix using weightings for the rows and for the columns.

SVD is a powerful way to extract the structure of a matrix. The rationale is to summarise the relationships between the K variables by using a small number S of latent continuous variables. These latent variables, called the *principal components*, are linear combinations of the initial variables. Note that the power of SVD explains why it is a well known tool to compress high dimensional data. Indeed we need S principal components only instead of K variables, allowing the use of a smaller memory size to store the relevant information contained in the data matrix.

More precisely, when performing MCA, the weighted SVD can be defined as follows. Let the diagonal matrix \mathbf{R} denote the weighting for the rows. The standard choice for \mathbf{R} is $\frac{1}{I}\mathbb{1}_I$, with $\mathbb{1}_I$ being the identity matrix of dimension I , which corresponds to a uniform weighting. The weighting for the columns is defined by the diagonal matrix $\frac{1}{K}\mathbf{D}_\Sigma^{-1}$ with dimensions $J \times J$ where $\mathbf{D}_\Sigma = \mathbf{diag}(p_1^{\mathbf{x}_1}, \dots, p_{q_1}^{\mathbf{x}_1}, \dots, p_1^{\mathbf{x}_K}, \dots, p_{q_K}^{\mathbf{x}_K})$ and $p_\ell^{\mathbf{x}_k}$ is the proportion of observations taking the category ℓ on the variable \mathbf{x}_k . Lastly, $\mathbf{M}_{I \times J}$ denotes the matrix where each row is equal to the vector of the means of each column of \mathbf{Z} (the disjunctive table corresponding to $\mathbf{X}_{I \times K}$). From these weightings, MCA consists in performing the SVD of the matrix triplet $(\mathbf{Z} - \mathbf{M}, \frac{1}{K}\mathbf{D}_\Sigma^{-1}, \mathbf{R})$ (Greenacre 1984) which is equivalent to writing $(\mathbf{Z} - \mathbf{M})$ as

$$\mathbf{Z} - \mathbf{M} = \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{V}^\top \quad (2)$$

where the columns of $\mathbf{U}_{I \times J}$ are the left singular vectors satisfying the relationship $\mathbf{U}^\top \mathbf{R} \mathbf{U} = \mathbb{1}_I$; the columns of $\mathbf{V}_{J \times J}$ are the right singular vectors satisfying the relationship $\mathbf{V}^\top \frac{1}{K}\mathbf{D}_\Sigma^{-1} \mathbf{V} = \mathbb{1}_J$ and $\mathbf{\Lambda}_{J \times J}^{1/2} = \mathbf{diag}(\lambda_1^{1/2}, \dots, \lambda_J^{1/2})$ is the diagonal matrix of the singular values.

The first S principal components are given by $\widehat{\mathbf{U}}_{I \times S} \widehat{\mathbf{\Lambda}}_{S \times S}^{1/2}$, the product between the first S columns of \mathbf{U} and the diagonal matrix $\mathbf{\Lambda}^{1/2}$ restricted to its S first elements. In the same way, the first S loadings are given by $\widehat{\mathbf{V}}_{J \times S}$, the first S columns of \mathbf{V} .

The weighting used for the columns provides several properties of MCA. First, it ensures that all the variables contribute in a same way to the analysis, independently of how many categories each of them has (Greenacre and Blasius

2006). It can be seen as the equivalent of scaling for continuous variables in PCA. Moreover, the principal components are the continuous variables which are the most linked to all the variables in the sense of the squared correlation ratio, i.e. the proportion of variance of the principal component explained by a categorical variable, or in other words, the ratio of the between-categories variability over the total variability. Lastly, rare categories are well considered by this weighting because dummy variables with a low proportion have a higher weight. It also allows individuals that take rare categories to be highlighted, even if they constitute a small part of the set of individuals only.

From Eq. (2), an estimate for \mathbf{Z} can be derived:

$$\widehat{\mathbf{Z}} = \widehat{\mathbf{U}}\widehat{\mathbf{\Lambda}}^{1/2}\widehat{\mathbf{V}}^T + \mathbf{M}. \tag{3}$$

$\widehat{\mathbf{Z}}$ is the best approximation of \mathbf{Z} with the constraint of rank S (Eckart–Young theorem Eckart and Young 1936) in the sense defined by the Hilbert–Schmidt norm

$$\|\mathbf{A}\|_{\frac{1}{K}\mathbf{D}_{\Sigma}^{-1}\otimes\mathbf{R}} = \sqrt{\text{tr}\left(\mathbf{A}\frac{1}{K}\mathbf{D}_{\Sigma}^{-1}\mathbf{A}^T\mathbf{R}\right)}.$$

Equation (3) is called the *reconstruction formula*. Note that, contrary to \mathbf{Z} , $\widehat{\mathbf{Z}}$ is a fuzzy disjunctive table in the sense that its cells are real numbers and not only zeros and ones as in a classic disjunctive table. However, the sum per variable is still equal to one (Tenenhaus and Young 1985). Most of the values are contained in the interval $[0, 1]$ or close to it because $\widehat{\mathbf{Z}}$ is as close as possible to \mathbf{Z} which contains only zeros and ones, but values out of this interval can occur.

Performing MCA requires $J - K$ parameters corresponding to the terms useful for the centering and the weighting of the categories, plus IS for the left singular vectors, minus S , because of centering constraint and minus $\frac{S(S+1)}{2}$ for the orthonormal constraint. The number of independent parameters for the right singular vectors $((J - K)S - S - \frac{S(S+1)}{2})$ are obtained in the same way. Thus, the total number of independent parameters for MCA is $J - K + S(I - 1 + (J - K) - S)$ (Josse and Husson 2011; Candès and Tao 2009). This number of parameters increases linearly with the number of cells in the data set.

3.2 Single imputation using MCA

Josse et al. (2012) proposed an iterative algorithm called “iterative MCA” to perform single imputation using MCA. The main steps of the algorithm are as follows:

1. initialization $\ell = 0$: recode \mathbf{X} as disjunctive table \mathbf{Z} , substitute missing values by initial values (the proportions) and calculate \mathbf{M}^0 and \mathbf{D}_{Σ}^0 on this completed data set.

2. step ℓ :

- (a) perform the MCA, in other words the SVD of $\left(\mathbf{Z}^{\ell-1} - \mathbf{M}^{\ell-1}, \frac{1}{K}\left(\mathbf{D}_{\Sigma}^{\ell-1}\right)^{-1}, \frac{1}{I}\mathbb{1}_I\right)$ to obtain $\widehat{\mathbf{U}}^{\ell}$, $\widehat{\mathbf{V}}^{\ell}$ and $\left(\widehat{\mathbf{\Lambda}}^{\ell}\right)^{1/2}$;
- (b) keep the S first dimensions and use the reconstruction formula (3) to compute the fitted matrix:

$$\widehat{\mathbf{Z}}_{I \times J}^{\ell} = \widehat{\mathbf{U}}_{I \times S}^{\ell}\left(\widehat{\mathbf{\Lambda}}_{S \times S}^{\ell}\right)^{1/2}\left(\widehat{\mathbf{V}}_{J \times S}^{\ell}\right)^T + \mathbf{M}_{I \times J}^{\ell-1}$$

and the new imputed data set becomes $\mathbf{Z}^{\ell} = \mathbf{W} * \mathbf{Z} + (\mathbb{1} - \mathbf{W}) * \widehat{\mathbf{Z}}^{\ell}$ with $*$ being the element-wise product, $\mathbb{1}_{I \times J}$ being a matrix with only ones and \mathbf{W} a weighting matrix where $w_{ij} = 0$ if z_{ij} is missing and $w_{ij} = 1$ otherwise. The observed values are the same but the missing ones are replaced by the fitted values;

- (c) from the new completed matrix \mathbf{Z}^{ℓ} , $\mathbf{D}_{\Sigma}^{\ell}$ and \mathbf{M}^{ℓ} are updated.

3. steps (2.a), (2.b) and (2.c) are repeated until the change in the imputed matrix falls below a predefined threshold $\sum_{ij}(\widehat{z}_{ij}^{\ell-1} - \widehat{z}_{ij}^{\ell})^2 \leq \varepsilon$, with ε equals to 10^{-6} for example.

The iterative MCA algorithm leads to an imputation of the disjunctive table, as well as to an estimate of the MCA parameters. Josse et al. (2012) showed that this method is powerful to obtain MCA parameters from incomplete data compared to other procedures described in (Van der Heijden and Escofier 2003). However, the iterative MCA is used to impute the data set and does not aim at performing MCA with missing values.

The algorithm can suffer from overfitting issues, when missing values are numerous, when the relationships between variables are weak, or when the number of observations is low. To overcome these issues, a regularized version of it has been proposed (Josse et al. 2012). The rationale is to remove the noise in order to avoid instabilities in the prediction by replacing the singular values $\left(\sqrt{\widehat{\lambda}_s^{\ell}}\right)_{1 \leq s \leq S}$ of step (2.b) by

$$\text{shrunk singular values} \left(\sqrt{\widehat{\lambda}_s^{\ell} - \frac{1}{J-K-S} \sum_{t=S+1}^{J-K} \lambda_t^{\ell}}\right)_{1 \leq s \leq S}.$$

In this way, singular values are shrunk with a greater amount of shrinkage for the smallest ones. Thus, the first dimensions of variability take a more significant part in the reconstruction of the data than the others. Assuming that the first dimensions of variability are made of information and noise, whereas the last ones are made of noise only, this behaviour is satisfactory. Note that regularization strategies of singular values motivates numerous researches for continuous data (Verbanck et al. 2013; Shabalina and Nobel 2013; Josse and Sardy 2015;

Gavish et al. 2014). Geometrically, the regularization makes the individuals closer to the center of gravity. Concerning the cells of $\hat{\mathbf{Z}}$, the regularization makes the values closer to the mean proportions, more stable and more often in the interval $[0, 1]$.

The regularized iterative MCA algorithm enables us to impute an incomplete disjunctive table but not an initial incomplete categorical data set. A strategy to go from the imputed disjunctive table to an imputed categorical data set is required. We also suggest the use of the coin flipping approach (Sect. 2.3). Let us note that for each set of dummy variables coding for one categorical variable, the sum per row is equal to one, even if it contains imputed values. Moreover, most of the imputed cells are in the interval $[0, 1]$ or are close to it. Consequently, modifications of these cells are rarely required.

3.3 MI using MCA

To perform MI using MCA, we need to reflect the uncertainty concerning the principal components and loadings. To do so, we use a non-parametric bootstrap approach based on the specificities of MCA. Indeed, as seen in Sect. 3.1, MCA enables us to assign a weight to each individual. This possibility to include a weight for the individual is very useful when the same rows of the data set occur several times. Instead of storing each replicate, a weight proportional to the number of occurrences of each row can be used, allowing the storage only of the rows that are different. Thus, a non-parametric bootstrap, such as the one used for the MI using the normal distribution, can easily be performed simply by modifying the weight of the individuals: if an individual does not belong to the bootstrap replicate, then its weight is null, otherwise, its weight is proportional to the number of times the observation occurs in the replicate. Note that individuals with a weight equal to zero are classically called *supplementary individuals* in the MCA framework (Greenacre 1984).

Thus, we define a MI method called multiple imputation using multiple correspondence analysis (MIMCA). First, the algorithm consists in drawing M sets of weights for the individuals. Then, for each set, a single imputation is performed: at first, the regularized iterative MCA algorithm is used to impute the incomplete disjunctive table using the given weights for the individuals; next, coin flipping is used to obtain categorical data and mimic the distribution of the categorical data. At the end, M imputed data sets are obtained and any statistical method can be applied on each one. In detail, the MIMCA algorithm is written as follows:

1. Reflect the variability on the set of parameters of the imputation model: draw I values with replacement in $\{1, \dots, I\}$ and define a weight r_i for each individual pro-

portional to the number of times the individual i is drawn. The weights are gathered in the weighting matrix \mathbf{R}^{boot} .

2. Impute the disjunctive table according to the previous weighting:

- (a) initialization $\ell = 0$: recode \mathbf{X} as a disjunctive table \mathbf{Z} , substitute missing values by initial values (the proportions) and calculate \mathbf{M}^0 and \mathbf{D}_Σ^0 on this completed data set.
- (b) step ℓ :
 - i. perform the SVD of

$$\left(\mathbf{Z}^{\ell-1} - \mathbf{M}^{\ell-1}, \frac{1}{K} \left(\mathbf{D}_\Sigma^{\ell-1} \right)^{-1}, \mathbf{R}^{boot} \right)$$

to obtain $\hat{\mathbf{U}}^\ell, \hat{\mathbf{V}}^\ell$ and $(\hat{\boldsymbol{\lambda}}^\ell)^{1/2}$;

- ii. keep the S first dimensions and compute the fitted matrix:

$$\hat{\mathbf{Z}}^\ell = \hat{\mathbf{U}}^\ell \left(\hat{\boldsymbol{\lambda}}_{shrunken}^\ell \right)^{1/2} \left(\hat{\mathbf{V}}^\ell \right)^\top + \mathbf{M}^{\ell-1}$$

where $(\hat{\boldsymbol{\lambda}}_{shrunken}^\ell)^{1/2}$ is the diagonal matrix containing the shrunk singular values and derive the new imputed data set $\mathbf{Z}^\ell = \mathbf{W} * \mathbf{Z} + (\mathbf{1} - \mathbf{W}) * \hat{\mathbf{Z}}^\ell$

- iii. from the new completed matrix $\mathbf{Z}^\ell, \mathbf{D}_\Sigma^\ell$ and \mathbf{M}^ℓ are updated.

- (c) step (2.b) is repeated until convergence.

3. Mimic the distribution of the categorical data set using coin flipping on \mathbf{Z}^ℓ :

- (a) if necessary, modify suitably the values of \mathbf{Z}^ℓ : negative values are replaced by zero, and values higher than one are replaced by one. Then, for each set of dummy variables coding for one categorical variable, scale in order to verify the constraint that the sum is equal to one.
- (b) for imputed cells coding for one missing value, draw one category according to a multinomial distribution.

4. Create M imputed data sets: for m from 1 to M alternate steps 1, 2 and 3.

3.4 Properties of the imputation method

MI using MCA is part of the family of joint modelling MI methods, which means that it avoids the runtime issues of FCS. Most of the properties of the MIMCA method are directly linked to MCA properties. MCA provides an efficient summary of the two-way associations between variables, as well as the similarities between individuals. The imputation benefits from these properties and provides an imputation model sufficiently complex to apply then an analysis model

focusing on two-way associations between variables, such as a main effects logistic regression model. In addition, because of the relatively small number of parameters required to perform MCA, the imputation method works well even if the number of individuals is small or if the number of variables is large. These properties have been highlighted in previous works on imputation using principal component methods (Audigier et al. 2014a, b) in comparison with MI using the normal distribution. Lastly, rare categories are well considered, and do not constitute an issue for the method.

Since these two methods, MIMCA and the multiple imputation with the normal distribution, reflect the uncertainty on their parameters using a bootstrap procedure, provide several imputations of the disjunctive table, and then use the same strategy to go from the disjunctive table to the categorical data set, they seem very close. However, due to the different imputation models used, the imputation of each disjunctive tables differ on many points and MIMCA is much more than an adaptation of multiple imputation using the normal distribution.

The first one is that the imputation of the disjunctive table by MCA is a deterministic imputation, replacing a missing value by the most plausible value given by the estimate of the principal components and the estimate of the loadings. Then, coin flipping is used to mimic the distribution of the categorical data. On the contrary, the multiple imputation based on the normal distribution uses stochastic regressions to impute the disjunctive table, that is to say, a Gaussian noise is added to the conditional expectation given by the observed values. Then, coin flipping is used, adding uncertainty a second time.

The second difference between the two methods is the covariance of the imputed values. Indeed, the matrix $\widehat{\mathbf{Z}}^\ell$ contains the reconstructed data by the iterative MCA algorithm and the product $\widehat{\mathbf{Z}}^{\ell\top} \widehat{\mathbf{Z}}^\ell$ provides the covariance matrix of this data. The rank of it is S . On the contrary, the rank of the covariance matrix used to perform imputation using the normal distribution is $J - K$ (because of the constraint that the sum is equal to one per variable). Consequently, the relationships between imputed variables are different.

The third difference is the number of estimated parameters. Indeed, the number of parameters is a major drawback for the normal distribution because it approximately increases with the square of the number of columns of the disjunctive table. Thus, the multiple imputation using the normal distribution cannot be used for a data set with a high number of categories. On the contrary, the imputation using MCA requires a number of parameters linearly dependent on the number of cells. This property is essential from a practical point of view because it makes it easy to impute data sets with a high number of categories.

The fourth difference is the way to impute the disjunctive table conditionally to the observed values. Imputation using

the normal distribution requires to inverse the covariance matrix. This can be a serious drawback for categorical data because collinearity between dummy variables frequently occurs. In this case the conditional distributions of the missing values do not exist. The issue can be overcome by adding a ridge term on the diagonal of the covariance matrix, but this term needs to be tuned. On the contrary, MCA imputes the disjunctive table using the reconstruction formula (Eq. 3) that does not require any matrix inversion. This is very appealing since the method can deal with data sets where the relationships between variables are strong.

4 Simulation study

As mentioned in the introduction, the aim of MI methods is to obtain an inference on a quantity of interest ψ . Here, we focus on the parameters of a logistic regression without interaction, which is a statistical method frequently used for categorical data. At first, we present how to make inference for the parameters from a multiply imputed data set. Then, we explain how we assess the quality of the inference built, that is to say, the quality of the MI methods. Finally, the MI methods presented in Sects. 2 and 3 are compared through a simulation study based on real data sets. It thus provides more realistic performances from a practical point of view. The code to reproduce all the simulations with the R software (R Core Team 2014), as well as the data sets used, are available on the webpage of the first author.

4.1 Inference from imputed data sets

Each MI method gives M imputed data sets as outputs. Then, the parameters of the analysis model (for instance the logistic regression) as well as their associated variance are estimated from each one. We denote $(\widehat{\psi}_m)_{1 \leq m \leq M}$ the set of the M estimates of the model's parameters and we denote $(\widehat{\text{Var}}(\widehat{\psi}_m))_{1 \leq m \leq M}$ the set of the M associated variances. These estimates have to be pooled to provide a unique estimate of ψ and of its variance using Rubin's rules (Rubin 1987).

This methodology is explained for a scalar quantity of interest ψ . The extension to a vector is straightforward, proceeding in the same way element by element. The estimate of ψ is simply given by the mean over the M estimates obtained from each imputed data set:

$$\widehat{\psi} = \frac{1}{M} \sum_{m=1}^M \widehat{\psi}_m, \quad (4)$$

while the estimate of the variance of $\widehat{\psi}$ is the sum of two terms:

$$\widehat{Var}(\hat{\psi}) = \frac{1}{M} \sum_{m=1}^M \widehat{Var}(\hat{\psi}_m) + \left(1 + \frac{1}{M}\right) \frac{1}{M-1} \sum_{m=1}^M (\hat{\psi}_m - \hat{\psi})^2. \quad (5)$$

The first term is the within-imputation variance, corresponding to the sampling variance. The second one is the between-imputation variance, corresponding to the variance due to missing values. The factor $(1 + \frac{1}{M})$ is due to the fact that $\hat{\psi}$ is estimated from a finite number of imputed tables.

Then, the 95 % confidence interval is calculated as:

$$\hat{\psi} \pm t_{v,.975} \sqrt{\widehat{Var}(\hat{\psi})}$$

where $t_{v,.975}$ is the .975 critical value of the Student's t -distribution with v degrees of freedom estimated as suggested by (Barnard and Rubin 1999).

4.2 Simulation design from real data sets

The validity of MI methods are often assessed by simulation (Van Buuren 2012, p. 47). We design a simulation study using real data sets to assess the quality of the MIMCA method. Each data set is considered as a population data and denoted \mathbf{X}_{pop} . The parameters of the logistic regression model are estimated from this population data and they are considered as the true coefficients ψ . Then, a sample \mathbf{X} is drawn from the population. This step reflects the sampling variance. The values of the response variable of the logistic model are drawn according to the probabilities defined by ψ . Then, incomplete data are generated completely at random to reflect the variance due to missing values (Brand et al. 2003). The MI methods are applied and the inferences are performed. This procedure is repeated T times.

The performances of a MI method are measured according to three criteria (Van Buuren 2012, p. 47) the bias given by $\frac{1}{T} \sum_{t=1}^T (\hat{\psi}_t - \psi)$, the median (over the T simulations) of the confidence intervals width as well as the coverage. This latter is calculated as the percentage of cases where the true value ψ is within its 95 % confidence interval.

A coverage sufficiently close to the nominal level is required to consider that the inference is correct, but it is not sufficient, the confidence interval width should be as small as possible.

To appreciate the value of the bias and of the width of the confidence interval, it is useful to compare them to those obtained from two other methods. The first one consists in calculating the criteria for the data sets without missing values, which we named the “Full data” method. The second one is the listwise deletion. This consists in deleting the individuals with missing values. Because the estimates of

the parameters of the model are obtained from a subsample, the confidence intervals obtained should be larger than those obtained from multiple imputation.

4.3 Results

The methods described in this paper are performed using the following R packages: *missMDA* (Husson and Josse 2015) for MIMCA, *cat* (Harding et al. 2012) for MI using the saturated loglinear model, *Amelia* (Honaker et al. 2014, 2011) for MI using a normal distribution, *mi* (Gelman et al. 2013) for MI using the DPMPM method, *mice* (Van Buuren and Groothuis-Oudshoorn 2014; Buuren and Groothuis-Oudshoorn 2011) for the FCS approach using iterated logistic regressions and random forests. This latter package will also be used to pool the results from the imputed data sets. The tuning parameters of each MIMCA competitors are chosen according to their default values implemented in the R packages. Firstly, the tuning parameter of the MIMCA method, that is to say, the number of components, is chosen to provide accurate inferences. Its choice will be discussed later in Sect. 4.3.3.

The MI methods are assessed in terms of the quality of the inference as well as the time consumed from data sets covering many situations. The data sets differ in terms of the number of individuals, the number of variables, the number of categories per variable, the relationships between variables.

The evaluation is based on the following categorical data sets. For each data set a categorical response variable is available.

- *Saheart*: This data set (Rousseauw et al. 1983) provides clinical attributes of $I_{pop} = 462$ males of the Western Cape in South Africa. These attributes can explain the presence of a coronary heart disease. The data set contains $K = 10$ variables with a number of categories between 2 and 4.
- *Galetas*: This data set (Applied Mathematics Department 2010) refers to the preferences of $I_{pop} = 1192$ judges regarding 11 cakes in terms of global appreciation and in terms of color aspect. The data set contains $K = 4$ variables with two that have 11 categories.
- *Sbp*: The $I_{pop} = 500$ subjects of this data set are described by clinical covariates explaining their blood pressure (GlaxoSmithKline 2003). The data set contains $K = 18$ variables that have 2 to 4 categories.
- *Income*: This data set, from the R package *kernlab* (Karatzoglou et al. 2004), contains $I_{pop} = 6876$ individuals described by several demographic attributes that could explain the annual income of an household. The data set contains $K = 14$ variables with a number of categories between 2 and 9.

Table 1 Bias over the several quantities of interest for several methods (Listwise deletion, Loglinear model, DPMPM, Normal distribution, MIMCA, FCS using logistic regressions, FCS using random forests, Full data) for the data set Titanic

	$\psi_1 = 2.25$	$\psi_2 = 0.86$	$\psi_3 = -0.16$	$\psi_4 = -0.92$	$\psi_5 = -1.06$	$\psi_6 = -2.42$
LD	0.74	-0.06	-0.14	-0.24	-0.41	-0.35
Loglinear	-0.10	-0.06	0.00	-0.04	0.16	-0.09
DPMPM	-0.77	-0.01	0.29	0.45	0.43	0.24
Normal	-0.86	-0.07	0.19	0.32	0.45	0.41
MIMCA 5	0.15	-0.01	-0.05	-0.15	0.00	-0.18
FCS-log	0.22	-0.06	-0.09	-0.19	-0.01	-0.22
FCS-forests	-1.24	0.00	0.33	0.57	0.47	0.73
Full data	0.10	-0.02	-0.03	-0.10	-0.01	-0.11

– *Titanic*: This data set (Dawson 1995) provides information on $I_{pop} = 2201$ passengers on the ocean liner *Titanic*. The $K = 4$ variables deal with the economic status, the sex, the age and the survival of the passengers. The first variable has four categories, while the other ones have two categories. The data set is available in the R software.

– *Credit*: German Credit Data from the UCI Repository of Machine Learning Database (Lichman 2013) contains $I_{pop} = 982$ clients described by several attributes which enable the bank to classify themselves as good or bad credit risk. The data set contains $K = 20$ variables with a number of categories between 2 and 4.

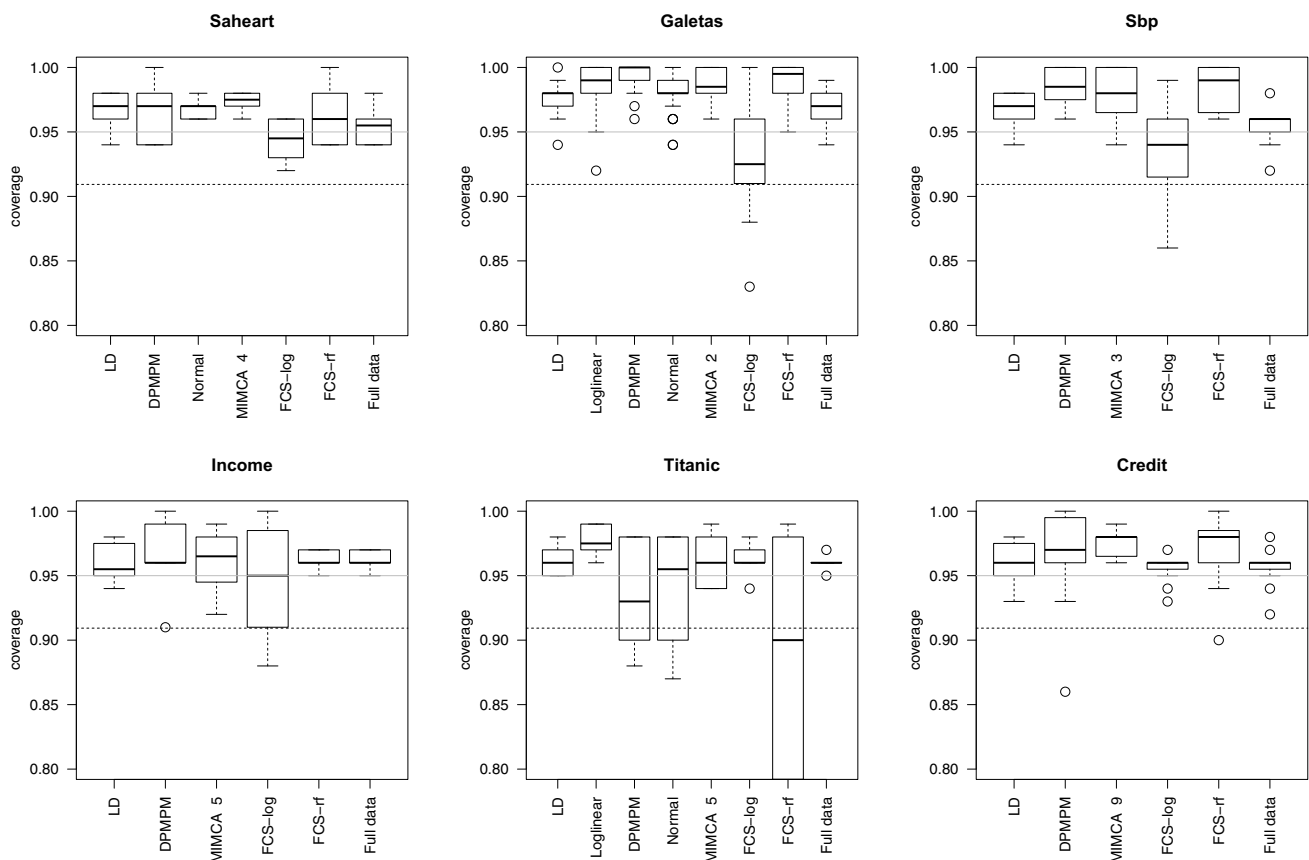


Fig. 1 Distribution of the coverages of the confidence intervals for all the parameters, for several methods (Listwise deletion, Loglinear model, Normal distribution, DPMPM, MIMCA, FCS using logistic regressions, FCS using random forests, Full data) and for different data sets (Saheart, Galetas, Sbp, Income, Titanic, Credit). The horizontal

dashed line corresponds to the lower bound of the 95 % confidence interval for a proportion of 0.95 from a sample of size 200 according to the Agresti-Coull method (Agresti and Coull 1998). Coverages under this line are considered as undesirable

The simulation design is performed for $T = 200$ simulations and 20 % of missing values generated independently and completely at random in each variable. The MI methods are performed with $M = 5$ imputed data sets which can be considered as satisfactory (Rubin 1987).

4.3.1 Assessment of the inferences

First of all, we can note that some methods cannot be applied on all the data sets. As explained previously, MI using the loglinear model can be applied only on data sets with a small number of categories such as *Titanic* or *Galetas*. MI using the normal distribution encounters inversion issues when the number of individuals is small compared to the number of categories. That is why no results are provided for MI using the normal distribution on the data sets *Credit* and *Sbp*. The others MI methods can be on all the data sets.

For each data set and each method, the coverages of all the confidence intervals of the parameters of the model are calculated from T simulations. Note that the number of parameters is between 6 and 30 (see Table 3 in Appendix 1 for more details on these models). All the coverages are summarized with a boxplot (see Fig. 1). The biases for one of the data set are provided in Table 1, similar results can be observed for the others (see Fig. 4 in Appendix). Exhaustive results for the confidence interval width are provided in Fig. 5 in Appendix.

As expected, MI using the loglinear model performs well on the two data sets where it can be performed. The coverages are close to the nominal levels, the biases are close to zero, and the confidence interval widths are small.

MI using the non-parametric version of the latent class model (DPMPM) performs quite well since most of the quantities of interest have a coverage close to 95 %. However, some inferences are incorrect from time to time such as on the data set *Credit* or *Titanic*. This behaviour is in agreement with the study of Si and Reiter (2013) which also presents some unsatisfactory coverages. Vidotto et al. (2014) note that this MI model may have some difficulties in capturing the associations among the variables, particularly when the number of variables is high or the relationships between variables are complex, that can explain the poor coverages observed. Indeed, on the data set *Credit*, the number of variables is the highest among the data sets considered, while on the data set *Titanic*, the relationships between variables can be described as complex, in the sense that the survival status of the passengers is linked to all the other variables, but these are not closely connected.

MI using the normal distribution can be applied on three data sets only. On these data sets, the coverages can be too small (see *Titanic* in Fig. 1). This highlights that despite the fact that this method is still often used in practice to deal with incomplete categorical data, it is not suitable and we do not

recommend using such a strategy. However, Schafer (1997) showed that this method could be used to impute mixed data (i.e. with continuous and categorical data) but only for complete categorical variables.

The FCS using logistic regressions encounters difficulties on the data sets with a high number of categories such as *Galetas* and *Income*. This high number of categories implies a high number of parameters for each conditional model that may explain the undercoverage on several quantities.

The FCS using random forests performs well except for the *Titanic* data set. The difficulties encountered for the data set can be explained by the step of subsampling variables in the imputation algorithm (Sect. 2.4). Indeed each tree is built with $\sqrt{K - 1}$ variables whereas the relationships between the variables are weak and all the variables are important to predict the survival response. Thus, it introduces too much bias in the individual tree prediction which may explain the poor inference. Even if, in the most practical cases, MI using random forests is very robust to the misspecification of the parameters, on this data set, the inference could be improved in increasing the number of explanatory variables retained for each tree.

Concerning MI using MCA, all the coverages observed are satisfying. The confidence interval width is of the same order of magnitude than the other MI methods. In addition, the method can be applied whatever the number of categories per variables, the number of variables or the number of individuals. Thus, it appears to be the easiest method to impute categorical data.

4.3.2 Computational efficiency

MI methods can be time consuming and the running time of the algorithms could be considered as an important property of a MI method from a practical point of view. Table 2 gathers the times required to impute $M = 5$ times the data sets with 20 % of missing values.

First of all, as expected, the FCS method is more time consuming than the others based on a joint model. In particular, for the data set *Income*, where the number of individuals and variables is high, the FCS using random forests requires 6329 seconds (i.e. 1.75 h), illustrating substantial running time issues. FCS using logistic regressions requires 881 seconds, a time 6 times higher than MI using the latent class model, and 15 times higher than MI method using MCA. Indeed, the number of incomplete variables increases the number of conditional models required, as well as the number of parameters in each of them because more covariates are used. In addition, the time required to estimate its parameters is non-negligible, particularly when the number of individuals is high. Then, MI using the latent class model can be performed in a reasonable time, but this is at least two times higher than the one required for MI using MCA. Thus, the MIMCA method

Table 2 Time consumed (in seconds) to impute data sets (Saheart, Galetas, Sbp, Income, Titanic, Credit), for different methods (Loglinear model, DPMPM, Normal distribution, MIMCA, FCS using logistic regressions, FCS using random forests)

	Saheart	Galetas	Sbp	Income	Titanic	Credit
Loglinear	NA	4.597	NA	NA	0.740	NA
DPMPM	20.050	17.414	56.302	143.652	10.854	24.289
Normal	0.920	0.822	NA	26.989	0.483	NA
MIMCA	5.014	8.972	7.181	58.729	2.750	8.507
FCS log	20.429	38.016	53.109	881.188	4.781	56.178
FCS forests	91.474	112.987	193.156	6329.514	265.771	461.248

The imputation is done for $M = 5$ data sets. Calculation has been performed on an Intel®Core™2 Duo CPU E7500, running Ubuntu 12.04 LTS equipped with 3 GB ram. Some values are not provided because all methods cannot be performed on each data set

should be particularly recommended to impute data sets of high dimensions.

Having a method which is not too expensive enables the user to produce more than the classical $M = 5$ imputed data sets. This leads to a more accurate inference.

4.3.3 Choice of the number of dimensions

MCA requires a predefined number of dimensions S which can be chosen by cross-validation (Josse et al. 2012). Cross-validation consists in selecting S which minimizes an error of prediction. More precisely, a small part of missing values is added completely at random to the data set \mathbf{X} (by default 5 %). Then, the missing values of the incomplete disjunctive table \mathbf{Z} are predicted using the regularized iterative MCA algorithm. The mean squared error of prediction is calculated according to $\frac{1}{\text{Card}(\mathcal{U})} \sum_{(i,j) \in \mathcal{U}} (z_{ij} - \hat{z}_{ij})^2$, where \mathcal{U} denotes the set of the added missing values. The procedure is repeated $nbsim$ times for each number of dimensions (by default $nbsim = 100$). The chosen number of dimensions is the one minimizing the mean of the $nbsim$ mean squared errors of prediction. This procedure of ‘repeated cross-validation’ can be used whether the data set contains missing values or not.

To assess how the choice of S impacts on the quality of the inference, we performed the simulations using number of dimensions closed to the one provided by the cross-validation procedure. Note that in the specific case of a simulation study from real data sets, the ‘true’ number of dimensions is unknown and its estimation by cross-validation may vary according to the incomplete sample drawn. For this reason, we used the most frequent number of dimensions obtained by cross-validation on ten incomplete data sets. Figure 2 presents how this tuning parameter influences the coverages in the previous study. The impacts on the width of the confidence intervals are reported in Fig. 6 and the ones on the bias in Fig. 7 in Appendix 2.

Except for the data set *Titanic*, the coverages are stable according to the chosen number of dimensions. In particular, the number of dimensions suggested by cross-validation provides coverages close to the nominal level of the con-

fidence interval. In the case of the data set *Titanic*, the cross-validation suggests retaining 5 dimensions, which is the choice giving the smallest confidence intervals, while giving coverages close to 95 %. But retaining less dimensions leads to worse performances since the covariates are not closely related (Sect. 4.3.1). Indeed, these covariates cannot be well represented within a space of low dimensions. Consequently, a high number of dimensions is required to reflect the useful associations to impute the data. *Titanic* illustrates that underfitting can be problematic. The same comment is made by Vermunt et al. (2008) who advise choosing a number of classes sufficiently high in the case of MI using the latent class model.

Note that cross-validation increases the time required to apply the algorithm. However, performing repeated cross-validation for a small number of times $nbsim$, the total computational time consumed remains small. For instance, $nbsim = 3$ is sufficient for the Income data set, adding 44 seconds to the MIMCA procedure. Thus, it remains the quicker strategy to impute categorical data.

5 Real data analysis

In this section we illustrate how to analyse an incomplete categorical data set using the *Titanic* data set (described in Sect. 4.3) as an example. We first detail how to tune the parameter of the method MIMCA. Then, we show how to perform multiple imputation step-by-step: generating multiple imputed data sets, applying a statistical method on each of them and combining the obtained results according to the Rubin’s rules. The MIMCA method is implemented in the R package missMDA (Husson and Josse 2015; Josse and Husson 2015).

5.1 Number of dimensions

To generate multiple imputed data sets with MIMCA, the first step consists in estimating the number of dimensions. Several cross-validation procedures are implemented in the

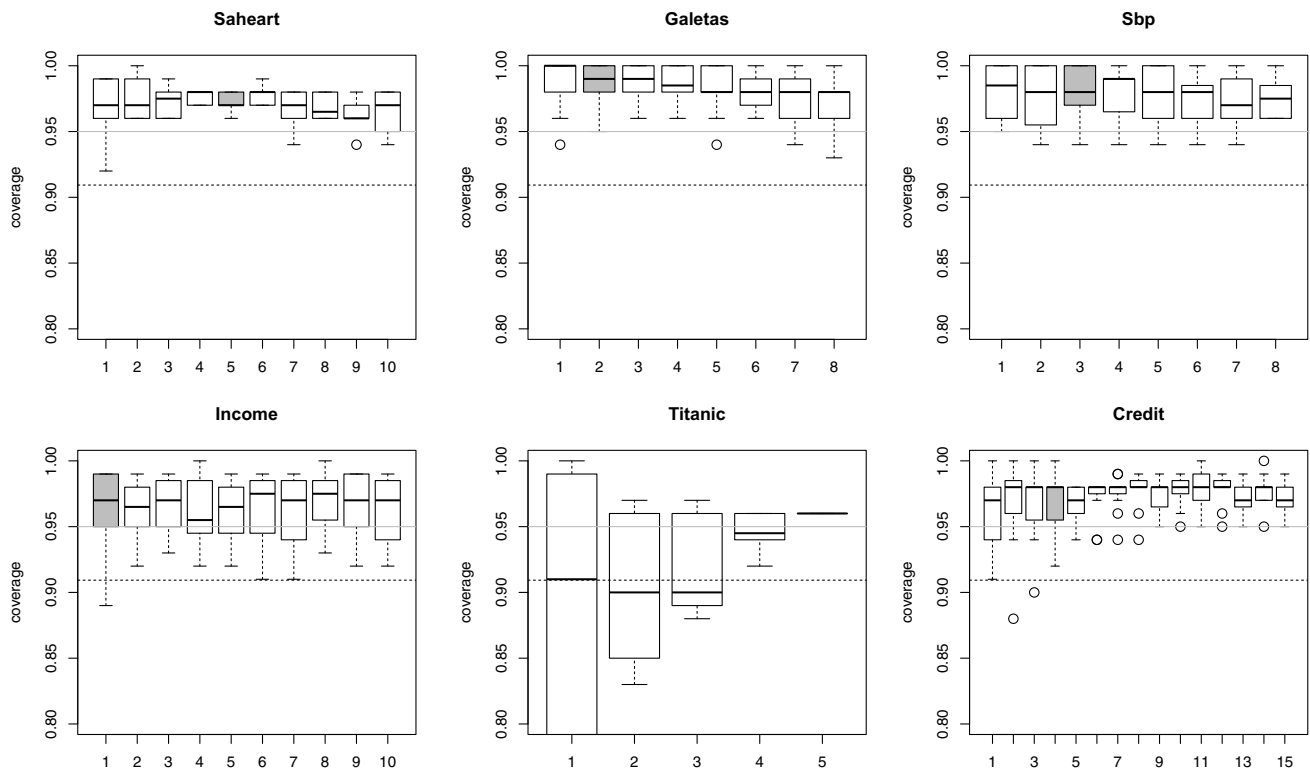


Fig. 2 Distribution of the coverages of the confidence intervals for all the parameters for the MIMCA algorithm for several numbers of dimensions and for different data sets (Saheart, Galetas, Sbp, Income, Titanic, Credit). The results for the number of dimensions provided by cross-validation are in grey. The horizontal dashed line corresponds to

the lower bound of the 95 % confidence interval for a proportion of 0.95 from a sample of size 200 according to the Agresti-Coull method (Agresti and Coull 1998). Coverages under this line are considered as undesirable

function `estim_ncpMCA`. The first one is leave-one-out cross-validation, which consists in predicting independently each observed cells of the data set with MCA. It can be run as follows:

```
> library(missMDA)
> data(TitanicNA)
> res.ncp <- estim_ncpMCA(TitanicNA,
  method.cv="loo")
```

The function outputs a list with the cross-validation error for each number of components, as well as the number of dimensions minimizing the cross-validation error. A progress bar indicates the speed to complete the run. If it is too slow, then the repeated cross-validation method detailed in Sect. 4.3.3 can be used. This method is less computationally intensive, but it is based on a simulation process which introduces a simulation error on the estimated number of dimensions. This method requires to specify additional parameters: the percentage of missing values added (`pNA`) and the number of missing data patterns generated (`nbsim`). Note that the percentage of missing values has to be small to preserve the data structure. By default, 5 % of missing values are added completely at random (`pNA=0.05`). The default number of missing data patterns simulated is `nbsim=100`.

A lower value makes the run faster, while a higher value decreases the simulation error. We recommend to run the method with the default parameters. Then, if the run is too slow, the process can be stopped and a smaller value of `nbsim` could be considered. Repeated cross-validation with twenty generated missing data patterns is performed as follows:

```
> res.ncp.kfold <- estim_ncpMCA(TitanicNA,
  + method.cv = "kfold", nbsim = 20)
```

The sensitivity of the results to the value of `nbsim` can be assessed with the plot of the cross-validation error with respect to the number of dimensions retained. This plot can be obtained as follows:

```
> plot(names(res.ncp.kfold$criterion),
  + res.ncp.kfold$criterion,
  + xlab="number of dimensions", + ylab=
  "cv error")
```

A clear trend indicates that the simulation error does not modify the number of dimensions minimizing the cross-validation error. Consequently, the choice for `nbsim` can be considered as suitable. Otherwise, if the curve is chaotic, then `nbsim` should be increased, even if the time required is

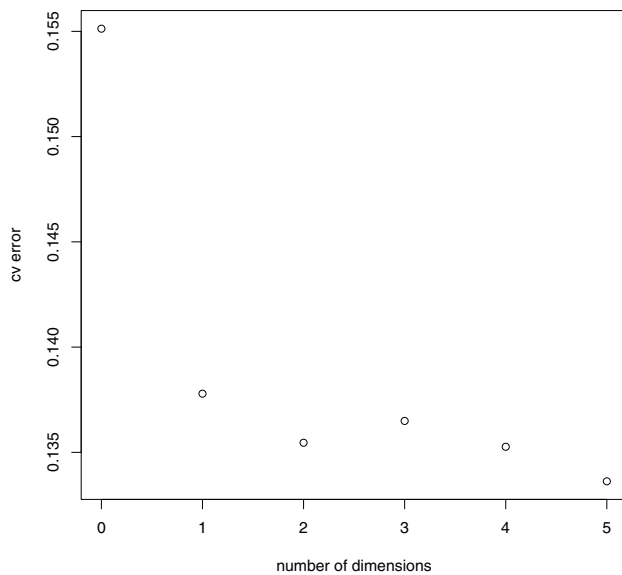


Fig. 3 Cross-validation error according to the number of dimensions used for the data set *Titanic*

important. For the data set *Titanic*, `nbsim=20` seems sufficient.

Figure 3 indicates to keep 5 dimensions for multiple imputation. Note that this number is provided in the object `res.ncp.kfold$ncp`.

5.2 Multiple imputation

We generate imputed data sets with the number of dimensions previously defined using the MIMCA function as follows:

```
> res.MIMCA <- MIMCA(TitanicNA,
+ ncp = res.ncp.kfold$ncp)
```

By default, `nboot=100` imputed data sets are generated.

To apply a statistical method on each imputed data set obtained from the MIMCA function and to combine the analysis results, we suggest using the R package *mice* (Van Buuren and Groothuis-Oudshoorn 2014).

For instance, to predict the survival status of the passengers of the ocean liner *Titanic* according to their age, sex and economic status, a logistic regression model can be applied as follows:

```
> # transform the MIMCA output as a
  mids object
> imp<-prelim(res.MIMCA,TitanicNA)
> # perform analysis
> fit <- with(data=imp,
+ exp=glm(SURV~CLASS+AGE+SEX,family =
+ "binomial"))
> # pool the analysis results
> summary(pool(fit))
```

The classical outputs with the coefficient estimates and their variance are obtained. Note that other R packages than *mice* can be used to make inference from the multiply imputed data set (e.g. the *BaBooN* package Meinfelder and Schnapp 2015).

6 Conclusion

This paper proposes an original MI method to deal with categorical data based on MCA. The principal components and the loadings that are the parameters of the MCA enable the imputation of data. To perform MI, the uncertainty on these parameters is reflected using a non-parametric bootstrap, which results in a specific weighting for the individuals.

From a simulation study based on real data sets, this MI method has been compared to the other main available MI methods for categorical variables. We highlighted the competitiveness of MIMCA to provide valid inferences for an analysis model requiring two-way associations (such as logistic regression without interaction, or loglinear model with two-way associations, proportion, odds ratios, etc).

We showed that MIMCA can be applied to various configurations of data. In particular, the method is accurate for a large number of variables, for a large number of categories per variable and when the number of individuals is small. Moreover, MIMCA runs fairly quickly, allowing the user to generate more imputed data sets and therefore to obtain more accurate inferences (Van Buuren 2012, p. 49 recommends to choose M between 20 and 100). Thus, MIMCA is very suitable to impute data sets of high dimensions that require more computation. Note that MIMCA depends on a tuning parameter (the number of components), but the performances of the MI method are robust to a misspecification of it.

The multiple imputation method based on MCA is implemented in the package *missMDA* (Husson and Josse 2015) of the R software. The function named MIMCA takes as input the incomplete data set, the number of dimensions used to impute the data, as well as the number of imputed tables M and returns a list with the imputed data sets.

Because of the intrinsic properties of MCA, MI using MCA is appropriate when the analysis model contains two-way associations between variables such as logistic regression without interaction. To consider the case with interactions, one solution could be to introduce to the data set additional variables corresponding to the interactions. However, the new variable “interaction” is considered as a variable in itself without taking into account its explicit link with the associated variables. It may lead to imputed values which are not in agreement with the others. This topic is a subject of intensive research for continuous variables (Seaman et al. 2012; Bartlett et al. 2014).

MIMCA is dedicated to perform multiple imputation for categorical data only. Another structure of data that is very

common is multilevel data. It should be interesting to investigate extension of MCA for such data. Note that a popular approach to handle such data is the method based on latent normal model suggested in (Carpenter and Kenward 2013) and implemented in the R package Jomo (Quartagno and Carpenter 2015) or the REALCOM-IMPUTE software (Carpenter et al. 2011). This method is also under investigation to handle unclustered categorical data.

In addition, MIMCA considered ordered categorical variables as nominal ones. Taking into account this specificity could improve the imputation and the subsequent inferences.

Lastly, the encouraging results of the MIMCA to impute categorical data prompt the extension of the method to impute mixed data. The first research in this direction (Audigier et al. 2014b) has shown that the principal component method dedicated to mixed data (called Factorial Analysis for Mixed Data) is efficient to perform single imputation, but the extension to a MI method requires further research.

Appendix 1: Simulation design: analysis models and sample characteristics

See Table 3.

Table 3 Set of the sample characteristics and of the analysis models used to perform the simulation study (Sect. 4.2) for the several data sets (Saheart, Galetas, Sbp, Income, Titanic, Credit)

Data set	Number of individuals	Number of variables	Sample size	Logistic regression model	Number of quantities of interest
Saheart	462	10	300	CHD = FAMHIST + TOBACCO + ALCOHOL	30
Galetas	1192	4	300	GALLE = GRUPO	6
Sbp	500	18	200	SBP = SMOKE + EXERCISE + ALCOHOL	12
Income	6876	14	1500	INCOME = SEX	8
Titanic	2201	4	300	SURV = CLASS+AGE+SEX	6
Credit	982	20	300	CLASS = CHECKING_STATUS + DURATION + CREDIT_HISTORY + PURPOSE	11

Appendix 2: Simulation study: complementary results

See Figs. 4–7.

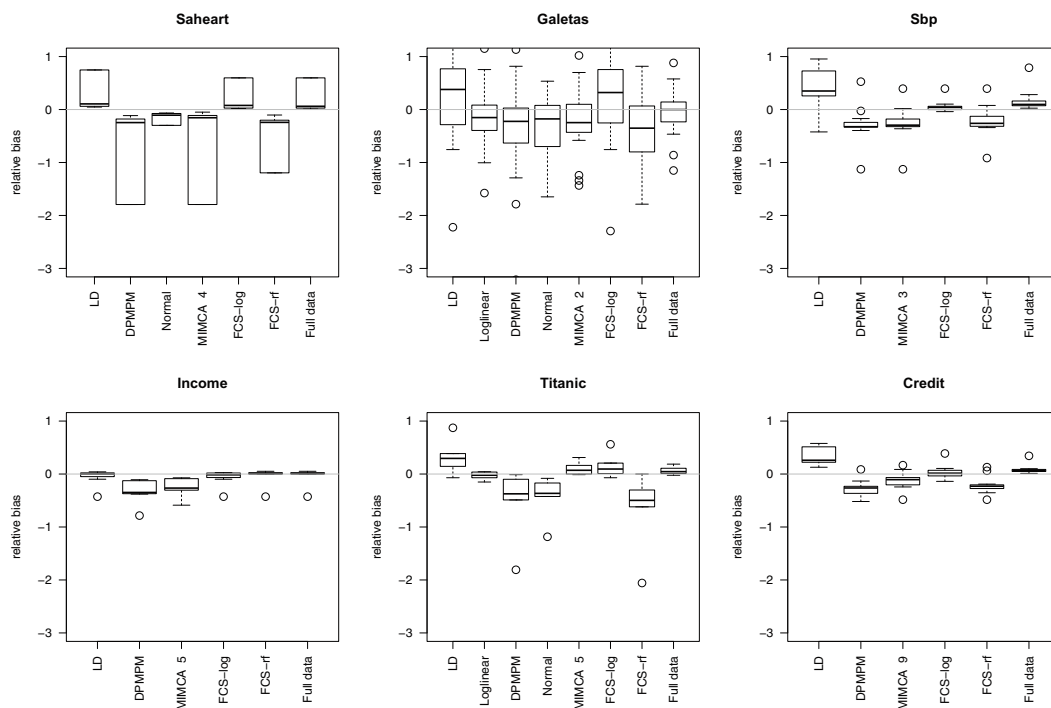


Fig. 4 Distribution of the relative bias (bias divided by the true value) over the several quantities of interest for several methods (Listwise deletion, Loglinear model, DPMPM, Normal distribution, MIMCA, FCS

using logistic regressions, FCS using random forests, Full data) for different data sets (Saheart, Galetas, Sbp, Income, Titanic, Credit). One point represents the relative bias observed for one coefficient

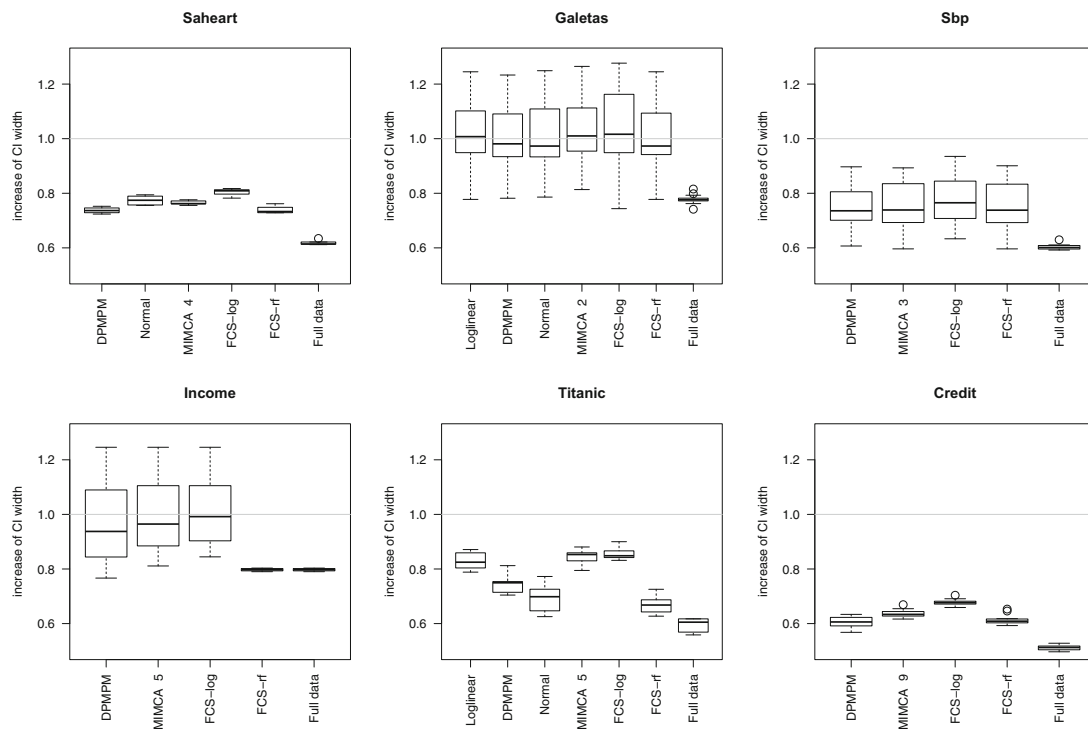


Fig. 5 Distribution of the median of the confidence interval for the several quantities of interest for several methods (Loglinear model, DPMPM, Normal distribution, MIMCA, FCS using logistic regressions, FCS using random forests, Full data) for different data sets (Saheart, Galetas, Sbp, Income, Titanic, Credit). One point represents the median

of the confidence interval observed for one coefficient divided by the one obtained by Listwise deletion. The *horizontal dashed line* corresponds to a ratio of 1. Points over this line corresponds to confidence interval higher than the one obtain by listwise deletion

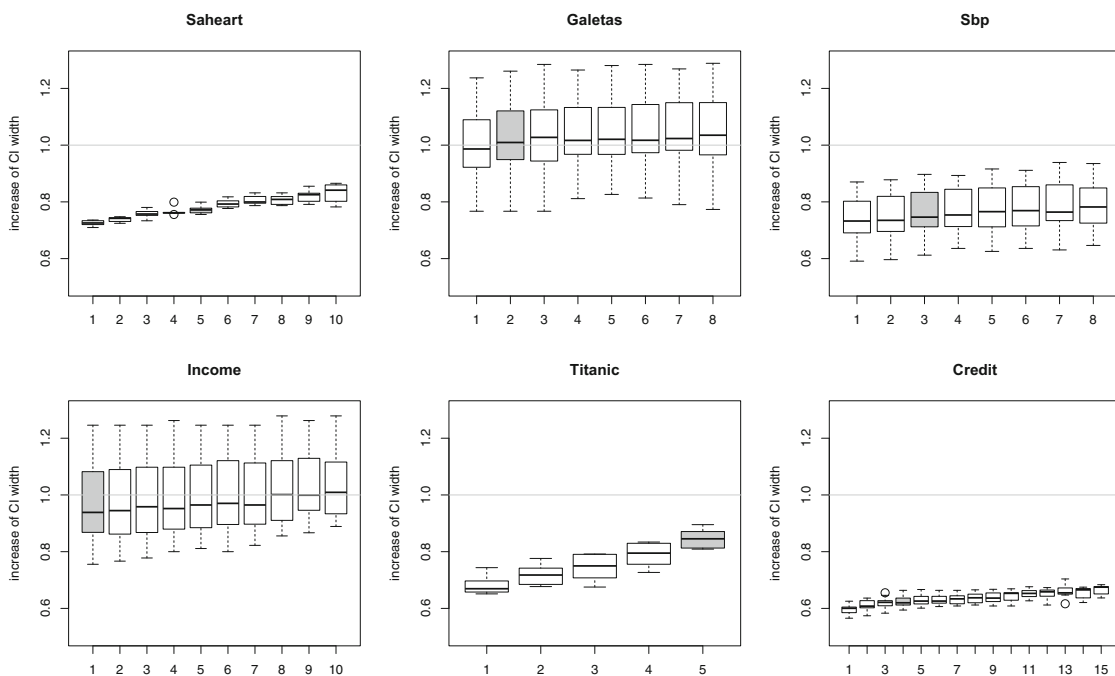


Fig. 6 Distribution of the median of the confidence interval for the several quantities of interest for the MIMCA algorithm for several numbers of dimensions for different data sets (Saheart, Galetas, Sbp, Income, Titanic, Credit). One point represents the median of the confidence interval observed for one coefficient divided by the one obtained by

Listwise deletion. The *horizontal dashed line* corresponds to a ratio of 1. Points over this line corresponds to confidence interval higher than the one obtain by listwise deletion. The results for the number of dimensions provided by cross-validation are in *grey*

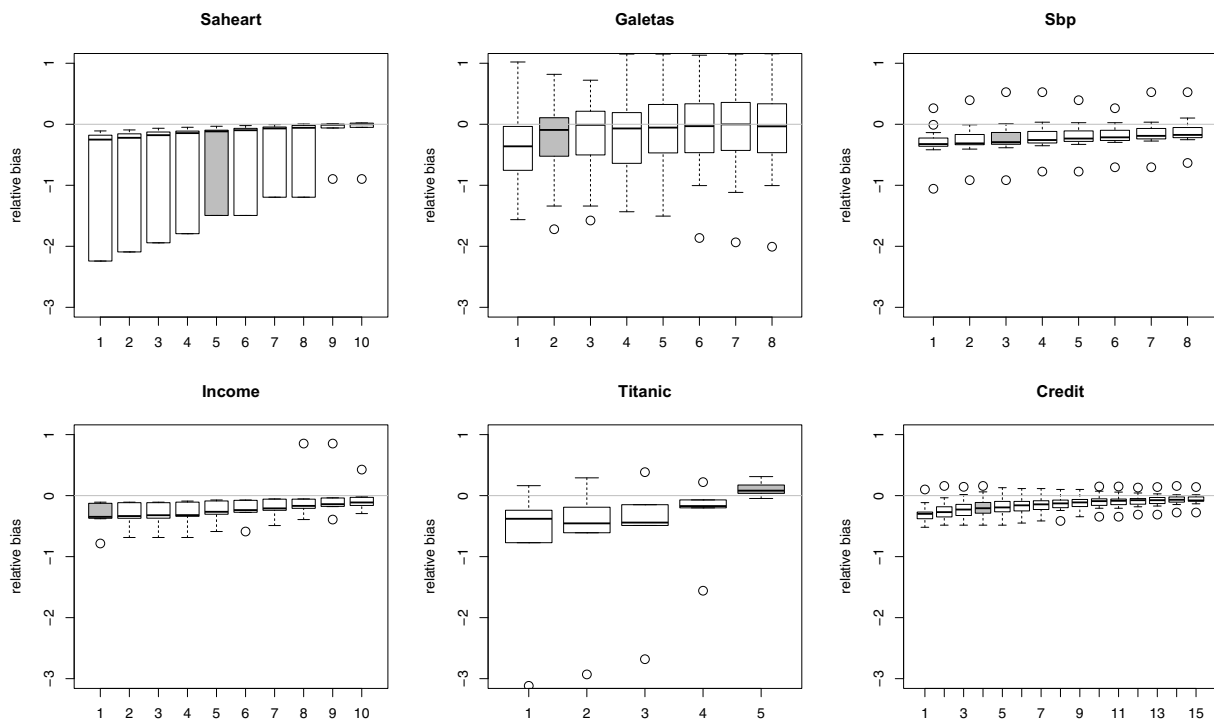


Fig. 7 Distribution of the relative bias (bias divided by the true value) over the several quantities of interest for the MIMCA algorithm for several numbers of dimensions for different data sets (Saheart, Gale-

tas, Sbp, Income, Titanic, Credit). One point represents the relative bias observed for one coefficient. The results for the number of dimensions provided by cross-validation are in grey

References

- Agresti, A.: Categorical Data Analysis. Wiley Series in Probability and Statistics. Wiley, New York (2013)
- Agresti, A., Coull, B.A.: Approximate is better than ‘exact’ for interval estimation of binomial proportions. *Am. Stat.* **52**(2), 119–126 (1998). doi:[10.2307/2685469](https://doi.org/10.2307/2685469)
- Albert, A., Anderson, J.A.: On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **71**(1), 1–10 (1984). doi:[10.2307/2336390](https://doi.org/10.2307/2336390)
- Allison, P.D.: Handling missing data by maximum likelihood. In: SAS global forum, pp 1–21 (2012)
- Allison, P.D.: Missing Data. Sage, Thousand Oaks (2002)
- Applied Mathematics Department, Agrocampus O, France (2010) galetas data set. http://math.agrocampus-ouest.fr/infogluce/DeliverLive/digitalAssets/74258_galetas.txt
- Audigier, V., Husson, F., Josse, J.: Multiple imputation for continuous variables using a Bayesian principal component analysis. *J. Stat. Comput. Simul.* (2014). doi:[10.1080/00949655.2015.1104683](https://doi.org/10.1080/00949655.2015.1104683)
- Audigier, V., Husson, F., Josse, J.: A principal component method to impute missing values for mixed data. *Adv. Data Anal. Classif.* **7**, 1–22 (2014)
- Barnard, J., Rubin, D.B.: Small sample degrees of freedom with multiple imputation. *Biometrika* **86**, 948–955 (1999)
- Bartlett, J.W., Seaman, S.R., White, I.R., Carpenter, J.R.: Multiple imputation of covariates by fully conditional specification: accommodating the substantive model. *Stat. Methods. Med. Res.* **24**, 462 (2014)
- Benzécri, J.P.: L’analyse des données. L’analyse des données. Tome II: L’analyse des correspondances. Dunod (1973)
- Bernaards, C.A., Belin, T.R., Schafer, J.L.: Robustness of a multivariate normal approximation for imputation of incomplete binary data. *Stat. Med.* **26**(6), 1368–1382 (2007)
- Besag, J.: Spatial interaction and the statistical analysis of lattice systems. *J. R. Stat. Soc. Ser. B (Methodological)* **36**(2), 192 (1974)
- Brand, J.P.L., van Buuren, S., Groothuis-Oudshoorn, K., Gelsema, E.S.: A toolkit in sas for the evaluation of multiple imputation methods. *Stat. Neerl.* **57**(1), 36–45 (2003). doi:[10.1111/1467-9574.00219](https://doi.org/10.1111/1467-9574.00219)
- Candès, E.J., Tao, T.: The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inf. Theory* **56**(5), 2053–2080 (2009). doi:[10.1109/TIT.2010.2044061](https://doi.org/10.1109/TIT.2010.2044061)
- Carpenter, J.R., Goldstein, H., Kenward, M.G.: REALCOM-IMPUTE software for multilevel multiple imputation with mixed response types. *J. Stat. Softw.* **45**(5), 1–14 (2011). <http://www.jstatsoft.org/v45/i05>
- Carpenter, J., Kenward, M.: Multiple Imputation and its Application, 1st edn. Wiley, Chichester (2013)
- Dawson, R.J.M.: The ‘unusual episode’ data revisited. *Journal of Statistics Education* **3**, 1–7. <http://www.amstat.org/publications/jse/v3n3/datasets.dawson.html> (1995)
- Demirtas, H.: Rounding strategies for multiply imputed binary data. *Biom. J.* **51**(4), 677–688 (2009)
- Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. *J. R. Stat. Soc. B* **39**, 1–38 (1977)
- Doove, L.L., Van Buuren, S., Dusseldorp, E.: Recursive partitioning for missing data imputation in the presence of interaction effects. *Comput. Stat. Data Anal.* **72**, 92–104 (2014). doi:[10.1016/j.csda.2013.10.025](https://doi.org/10.1016/j.csda.2013.10.025)
- Dunson, D.B., Xing, C.: Nonparametric Bayes modeling of multivariate categorical data. *J. Am. Stat. Assoc.* **104**(487), 1042–1051 (2009)
- Eckart, C., Young, G.: The approximation of one matrix by another of lower rank. *Psychometrika* **1**(3), 211–218 (1936)
- Gavish, M., Donoho, D.: Optimal shrinkage of singular values. [arXiv:1405.7511](https://arxiv.org/abs/1405.7511) e-prints (214)

- Gelman, A., Hill, J., Su, Y., Yajima, M., Grazia Pittau, M., Goodrich, B., Si, Y.: *mi*: Missing data imputation and model checking. R package version 0.9-93 (2013)
- Gifi, A.: *Nonlinear Multivariate Analysis*. D.S.W.O. Press, Leiden (1981)
- GlaxoSmithKline, Toronto, Ontario, Canada: Blood pressure data set. <http://www.math.yorku.ca/Who/Faculty/Ng/ssc2003/BPMainF.htm> (2003)
- Greenacre, M.J.: *Theory and Applications of Correspondence Analysis*. Academic Press, London (1984)
- Greenacre, M.J., Blasius, J.: *Multiple Correspondence Analysis and Related Methods*. Chapman & Hall/CRC, Boca Raton (2006)
- Harding, T., Tusell, F., Schafer, J.L.: *cat*: Analysis of categorical-variable datasets with missing values. <http://CRAN.R-project.org/package=cat>, r package version 0.0-6.5 (2012)
- Honaker, J., King, G., Blackwell, M.: *Amelia II*: A program for missing data. R package version 1.7.2 (2014)
- Honaker, J., King, G., Blackwell, M.: *Amelia II*: A program for missing data. *J. Stat. Softw.* **45**(7), 1–47 (2011)
- Husson, F., Josse, J.: *missMDA*: Handling missing values with multivariate data analysis. <http://CRAN.R-project.org/package=missMDA>, r package version 1.9 (2015)
- Ishwaran, H., James, L.: Gibbs sampling methods for stick-breaking priors. *J. Am. Stat. Assoc.* **96**(453), 161–173 (2001)
- Josse, J., Chavent, M., Liquet, B., Husson, F.: Handling missing values with regularized iterative multiple correspondence analysis. *J. Classif.* **29**, 91–116 (2012)
- Josse, J., Husson, F.: Selecting the number of components in PCA using cross-validation approximations. *Comput. Stat. Data Anal.* **56**(6), 1869–1879 (2011)
- Josse, J., Husson, F.: *missmda* a package to handle missing values in and with multivariate data analysis methods. *J. Stat. Softw.* **25**, 1 (2015)
- Josse, J., Sardy, S.: Adaptive shrinkage of singular values. *Stat. Comput.* **71**, 1–10 (2015)
- Karatzoglou, A., Smola, A., Hornik, K., Zeileis, A.: *kernlab*—an S4 package for kernel methods in R. *J. Stat. Softw.* **11**(9):1–20, <http://www.jstatsoft.org/v11/i09/> (2004)
- King, G., Honaker, J., Joseph, A., Scheve, K.: Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *Am. Polit. Sci. Rev.* **95**(1), 49–69 (2001)
- Lebart, L., Morineau, A., Werwick, K.M.: *Multivariate Descriptive Statistical Analysis*. Wiley, New-York (1984)
- Lichman, M.: UCI machine learning repository. <http://archive.ics.uci.edu/ml> (2013)
- Little, R.J.A., Rubin, D.B.: *Statistical analysis with missing data*. Wiley series in probability and statistics, Wiley, New-York (1987, 2002)
- Meinfielder, F., Schnapp, T.: *BaBooN*: Bayesian bootstrap predictive mean matching—multiple and single imputation for discrete data. <https://CRAN.R-project.org/package=BaBooN>, r package version 0.2-0 (2015)
- Meng, X.L., Rubin, D.B.: Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *J. Am. Stat. Assoc.* **86**(416), 899–909 (1991)
- Nishisato, S.: *Analysis of Categorical Data: Dual Scaling and its Applications*. University of Toronto Press, Toronto (1980)
- Quartagno, M., Carpenter, J.: *jomo*: A package for multilevel joint modelling multiple imputation. <http://CRAN.R-project.org/package=jomo> (2015)
- R Core Team: *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, <http://www.R-project.org/> (2014)
- Rousseauw, J., du Plessis, J., Benade, A., Jordann, P., Kotze, J., Jooste, P., Ferreira, J.: Coronary risk factor screening in three rural communities. *S. Afr. Med. J.* **64**, 430–436 (1983)
- Rubin, D.B.: *Multiple Imputation for Non-Response in Survey*. Wiley, New York (1987)
- Schafer, J.L.: *Analysis of Incomplete Multivariate Data*. Chapman & Hall/CRC, London (1997)
- Schafer, J.L.: Multiple imputation in multivariate problems when the imputation and analysis models differ. *Stat. Neerl.* **57**(1), 19–35 (2003)
- Seaman, S.R., Bartlett, J.W., White, I.R.: Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods. *BMC Med. Res. Methodol.* **12**(1), 46 (2012). doi:[10.1186/1471-2288-12-46](https://doi.org/10.1186/1471-2288-12-46)
- Shabalin, A., Nobel, B.: Reconstruction of a low-rank matrix in the presence of gaussian noise. *J. Multivar. Anal.* **118**, 67–76 (2013)
- Shah, A.D., Bartlett, J.W., Carpenter, J., Nicholas, O., Hemingway, H.: Comparison of random forest and parametric imputation models for imputing missing data using MICE: A CALIBER study. *Am. J. Epidemiol.* **179**(6), 764–774 (2014). doi:[10.1093/aje/kwt312](https://doi.org/10.1093/aje/kwt312)
- Si, Y., Reiter, J.: Nonparametric bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *J. Educ. Behav. Stat.* **38**, 499–521 (2013)
- Stekhoven, D.J., Bühlmann, P.: Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* **28**(1), 112–118 (2012)
- Tenenhaus, M., Young, F.W.: An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika* **50**, 91–119 (1985)
- Van Buuren, S., Groothuis-Oudshoorn, K.: *mice*. R package version 2.22 (2014)
- Van Buuren, S., Brand, J.P.L., Groothuis-Oudshoorn, C.G.M., Rubin, D.B.: Fully conditional specification in multivariate imputation. *J. Stat. Comput. Simul.* **76**, 1049–1064 (2006)
- Van Buuren, S.: *Flexible Imputation of Missing Data* (Chapman & Hall/CRC Interdisciplinary Statistics), 1st edn. Chapman and Hall/CRC, Boca Raton (2012)
- Van Buuren, S., Groothuis-Oudshoorn, C.G.M.: *mice*: Multivariate imputation by chained equations in R. *J. Stat. Softw.* **45**(3), 1–67 (2011)
- Van der Heijden, P., Escofier, B.: *Analyse des correspondances: recherches au coeur de l'analyse des données*, Presses universitaires de Rennes, Rennes, France, chap Multiple correspondence analysis with missing data, pp 152–170 (2003)
- van der Palm, D., van der Ark, L., Vermunt, J.: A comparison of incomplete-data methods for categorical data. *Stat. Methods Med. Res.* **17**, 33 (2014)
- Verbanck, M., Josse, J., Husson, F.: Regularised PCA to denoise and visualise data. *Stat. Comput.* **25**(2), 471–486 (2013). doi:[10.1007/s11222-013-9444-y](https://doi.org/10.1007/s11222-013-9444-y)
- Vermunt, J.K., van Ginkel, J.R., van der Ark, L.A., Sijtsma, K.: Multiple imputation of incomplete categorical data using latent class analysis. *Sociol. Methodol.* **38**(38), 369–397 (2008)
- Vidotto, D., Kapteijn, M.C., Vermunt, J.: Multiple imputation of missing categorical data using latent class models: State of art. *Psychol. Test Assess. Model.* **57**, 542 (2014)
- Yucel, R.M., He, Y., Zaslavsky, A.M.: Using calibration to improve rounding in imputation. *Am. Stat.* **62**, 125–129 (2008)